Received: 7 February 2017 | Revised: 11 December 2017 | Accepted: 20 December 2017

DOI: 10.1002/mpr.1608

ORIGINAL ARTICLE

WILEY

A tutorial on conducting genome-wide association studies: Quality control and statistical analysis

Andries T. Marees[1,2,3,4,5] | Hilde de Kluiver[6] | Sven Stringer[7] | Florence Vorspan[1,2,3,4,8,9] | Emmanuel Curis[3,10,11] | Cynthia Marie-Claire[2,3,4] | Eske M. Derks[1,5]

[1] Department of Psychiatry, Amsterdam Medical Center, Amsterdam, The Netherlands

[2] Inserm, UMR-S 1144, Paris, France

[3] Université Paris Descartes, UMR-S 1144, Paris, France

[4] Université Paris Diderot, Sorbonne Paris Cité, UMR-S 1144, Paris, France

[5] QIMR Berghofer, Translational Neurogenomics Group, Brisbane, Australia

[6] GGZ inGeest and Department of Psychiatry, Amsterdam Public Health research institute, VU University Medical Center, Amsterdam, The Netherlands

[7] Department of Complex Trait Genetics, VU University, Amsterdam, The Netherlands

[8] Service de Médecine Addictologique, APHP, Hôpital Fernand Widal, Paris, France

[9] Faculté de Médecine, Université Paris Diderot, Paris, France

[10] Laboratoire de biomathématiques, faculté de pharmacie de Paris, Université Paris Descartes, Paris, France

[11] Service de biostatistiques et informatique médicales, Hôpital Saint-Louis, APHP, Paris, France

Correspondence
Andries T. Marees, Department of Psychiatry, Academic Medical Center, Room PA1-179 Meibergdreef 5, 1105 AZ, Amsterdam, The Netherlands.
Email: a.t.marees@amc.uva.nl

Funding information
ANR, Grant/Award Number: ANR-11-IDEX-0004-02; Investissement d'Avenir programme; Foundation Volksbond Rotterdam

Abstract

Objectives: Genome-wide association studies (GWAS) have become increasingly popular to identify associations between single nucleotide polymorphisms (SNPs) and phenotypic traits. The GWAS method is commonly applied within the social sciences. However, statistical analyses will need to be carefully conducted and the use of dedicated genetics software will be required. This tutorial aims to provide a guideline for conducting genetic analyses.

Methods: We discuss and explain key concepts and illustrate how to conduct GWAS using example scripts provided through GitHub (https://github.com/MareesAT/GWA_tutorial/). In addition to the illustration of standard GWAS, we will also show how to apply polygenic risk score (PRS) analysis. PRS does not aim to identify individual SNPs but aggregates information from SNPs across the genome in order to provide individual-level scores of genetic risk.

Results: The simulated data and scripts that will be illustrated in the current tutorial provide hands-on practice with genetic analyses. The scripts are based on PLINK, PRSice, and R, which are commonly used, freely available software tools that are accessible for novice users.

Conclusions: By providing theoretical background and hands-on experience, we aim to make GWAS more accessible to researchers without formal training in the field.

KEYWORDS

genome-wide association study (GWAS), GitHub, PLINK, polygenic risk score (PRS), tutorial

1 | INTRODUCTION

During the past two decades, there has been a growing interest in investigating the influence of genetic risk factors on variation in human behaviour. The technical and analytic tools needed to conduct genetic studies have become increasingly accessible. This increased accessibility offers great promise as researchers outside the field of genetics may bring new expertise to the field (e.g., more in-depth knowledge of the nosology of psychiatric traits). However, performing genetic association studies in a correct manner requires specific

knowledge of genetics, statistics, and (bio)informatics. This paper aims to provide a guideline for conducting genetic analyses by introducing key concepts and by sharing scripts that can be used for data analysis.

The aim of genome-wide association studies (GWAS) is to identify **single nucleotide polymorphisms** (**SNPs**; see Box 1 for an explanation of all terms that are printed in bold throughout the manuscript) of which the allele frequencies vary systematically as a function of phenotypic trait values (e.g., between cases with schizophrenia and healthy controls, or between individuals with high vs. low scores on neuroticism). Identification of trait-associated SNPs may subsequently reveal new insights into the biological mechanisms underlying these phenotypes. Technological advancements allow investigation of the impact of large numbers of SNPs distributed throughout the genome.

To date, GWAS have been successful in revealing SNPs that contribute to the risk of psychiatric traits, including schizophrenia, autism spectrum disorders, attention deficit hyperactivity disorder, major depressive disorder, and bipolar disorder (Gelernter et al., 2014; Ripke et al., 2014; Smoller, 2013; Sullivan, Daly, & O'Donovan, 2012). The overall picture of these results suggest that psychiatric traits are influenced by many common as well as rare SNPs each having small individual effect sizes (Gibson, 2012). The aforementioned GWAS relied strongly on in-depth knowledge of the genetic architecture of the human genome, which was provided by two important research initiatives, namely, the International HapMap Project and the 1000 Genomes project. The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/; Gibbs et al., 2003) described the patterns of common SNPs within the human DNA sequence whereas the 1000 Genomes (1KG) project (http://www.1000genomes.org/; Altshuler et al., 2012) provided a map of both common and rare SNPs.

Because GWAS results showed that effect sizes of individual SNPs are small, researchers in the psychiatric field developed an interest in methods that aggregate the effect of SNPs. We will specifically focus on polygenic risk score (PRS) analysis as we believe this to be the most relevant method to present here, as it is relatively easy to conduct while it can be applied to target samples with relatively modest sample sizes (Dudbridge, 2013). PRS combines the effect sizes of multiple SNPs into a single aggregated score that can be used to predict disease risk (Dudbridge, 2016). The PRS is an individual-level score that is calculated based on the number of risk variants that a person carries, weighted by SNP effect sizes that are derived from an independent large-scaled discovery GWAS. As such, the score is an indication of the total genetic risk of a specific individual for a particular trait, which can be used for clinical prediction or screening (e.g., breast cancer; Shieh et al., 2016). For psychiatric traits, PRS is also significantly associated with case–control status; however, its discriminative accuracy is not (yet) sufficient for clinical applications (Vassos et al., 2017; Wray et al., 2013). PRS has contributed to our knowledge of the genetic architecture of psychiatric traits by its ability to predict disease status. It has further been used to investigate whether genetic effect sizes obtained from a GWAS of a specific phenotype of interest can be used to predict the risk of another phenotype (Derks et al., 2012; Ruderfer et al., 2014; Smoller, 2013; Stringer, Kahn, de Witte, Ophoff, & Derks, 2014).

Even though recent GWAS have led to the identification of significantly associated SNPs for many phenotypic traits, the contribution of social scientists and clinicians to the genetics field can improve our understanding of the precise nature of the specific behavioural, cognitive, or neural correlates of identified risk SNPs. However, the analysis of genetic data requires the execution of several quality checks and careful conductance of statistical analyses to avoid spurious associations due to several potential sources of confounding (e.g., ethnic stratification). In addition, at least a fair knowledge of genetic power calculation is necessary to avoid performing underpowered studies. For more information on how to conduct power analyses, we refer to another tutorial of our group (Stringer et al., 2015).

This tutorial provides a guideline to researchers who wish to incorporate genetics into their studies but do not have a formal background in this topic. First, we will show how to apply rigorous quality control (QC) procedures on genotype data prior to conducting GWAS, including the use of appropriate methods to take into account ethnic heterogeneity. Second, we illustrate commonly used tests of association between SNPs and phenotypic traits of interest while controlling for potential confounders. Third, we will show how to conduct PRS analysis. Example R and Unix scripts are provided at https://github.com/MareesAT/GWA_tutorial/. We do not elaborate on SNP imputation of unmeasured SNPs based on reference data. Although imputation is a commonly used method in GWAS, it is beyond the scope of this article. For a solid introduction on this topic, we refer to an article by van Leeuwen and colleagues (2015).

2 | SOFTWARE

QC procedures and statistical analyses will be illustrated using the free, open-source whole-genome association analysis toolset PLINK version 1.07 (Purcell et al., 2007) that can be downloaded from http://zzz.bwh.harvard.edu/plink/. The PLINK 1.9 beta version contains the same options, while being much faster https://www.cog-genomics.org/plink/1.9/. As PLINK 1.9 is currently a beta version, we have used the official PLINK version in this tutorial. However, it is also possible to complete all tutorials using PLINK 1.9. Even though some of the steps discussed in this article could be performed in conventional statistical packages such as R, a software package specifically dedicated to the analysis of genetic data is much more convenient to use. In addition to PLINK, there are many other good options available for the analysis of SNP data such as Genabel (Aulchenko, Ripke, Isaacs, & Van Duijn, 2007) and SNPTEST (Marchini, Howie, Myers, McVean, & Donnelly, 2007). Furthermore, methods that allow for testing association in family-based GWAS have also been developed (Chen & Yang, 2010; Ott, Kamatani, & Lathrop, 2011). We advise to use GNU/Linux-based computer resources although many of the options are also available through the windows version of PLINK. A basic introduction to shells and command lines can be found at http://www.ee.surrey.ac.uk/Teaching/Unix/. All graphs generated by the GitHub example scripts will be obtained using the free, open-source programming language R (https://www.r-project.org/).

2.1 | Data format

PLINK can either read text-format files or binary files. Because reading large text files can be time-consuming, it is recommended to use binary files. Text PLINK data consist of two files: one contains information on the individuals and their genotypes (*.ped); the other contains

information on the genetic markers (*.map; see Figure 1). In contrast, binary PLINK data consist of three files, a binary file that contains individual identifiers (IDs) and genotypes (*.bed), and two text files that contain information on the individuals (*.fam) and on the genetic markers (*.bim; see Figure 1). For example, in a study of bipolar disorder, the *.bed file would contain the genotyping results of all patients and healthy controls; the *.fam file would contain the subject-related data (family relationship with other participants in the study, sex, and clinical diagnosis); whereas the *.bim file would contain information on the physical position of the SNPs. Analysis using covariates often requires a fourth file, containing the values of these covariates for each individual (see Figure 1).

2.2 | Basic PLINK command

PLINK is a command line program; hence, its usage requires an active shell waiting for commands. This can be recognized by its prompt ($ or >) just before the cursor. Often, the path of the current directory will be displayed before the prompt, as in Figure 2. The current directory is a central notion for PLINK usage, because by default, PLINK will load data files from, and save result files in this directory. The current directory can be changed to any directory using conventional Unix commands, typically `cd`. After the prompt, the use of PLINK is indicated by typing the `plink` keyword. If PLINK is not installed in a standard directory, the path to the directory where PLINK is installed has to be typed in front of the command, for instance, `/usr/local/bin/plink`.

After the `plink` keyword, other options controlling PLINK's workflow will follow, separated by spaces. These options all begin with two dashes (--). One of the first options to provide is the format and the name of the data files: use `--file {your_file}` for text files and `--bfile {your_file}` for binary files. After that, all other required options can be added, for instance, the `--assoc` option to perform an association analysis as displayed in Figure 2; this specific option will tell PLINK to perform an X^2 test for each SNP to the phenotype of interest. Multiple options can be combined within a single command line. Within PLINK, a default order is implemented, which works regardless of the order of the commands within the command line. A useful, and sometimes mandatory, option is `--out {outfile}`, which provides a name to output files (suffixes will be added as needed by PLINK). Beware that PLINK will delete without notice any existent file with the same name. Please note that the options within PLINK extend beyond what is being discussed in the current article; for the full set of options, see http://zzz.bwh.harvard.edu/plink/.

3 | QC OF GENETIC DATA

A vital step that should be part of any GWAS is the use of appropriate QC. Without extensive QC, GWAS will not generate reliable results because raw genotype data are inherently imperfect. Errors in the data can arise for numerous reasons, for example, due to poor quality of DNA samples, poor DNA hybridization to the array, poorly performing genotype probes, and sample mix-ups or contamination. For instance, failing to thoroughly control for these data issues has led to the retraction of an article published by Sebastiani et al. (2010) in *Science* (Sebastiani et al., 2010, 2011; Sebastiani et al., 2012; Sebastiani et al., 2013). The results of the retracted article were affected by technical errors in the Illumina 610 array and an inadequate QC to account for

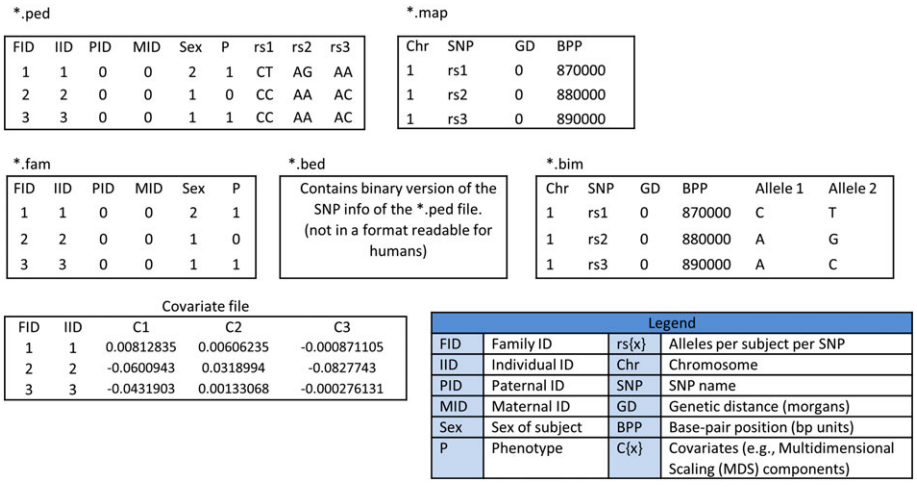

FIGURE 1 Overview of various commonly used PLINK files. SNP = single nucleotide polymorphism

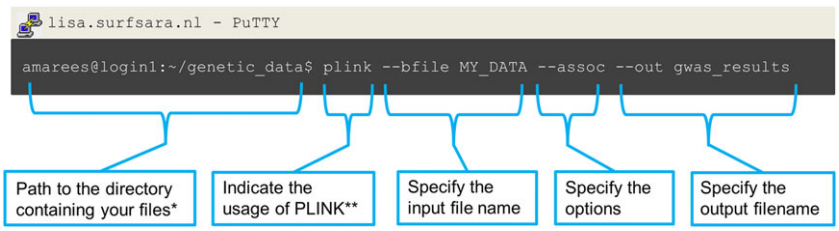

FIGURE 2 Structure of the PLINK command line. *Not all shells will show this. **Provide the path to the directory where PLINK is installed if this is not in the current directory (e.g., /usr/local/bin/plink). Note that this example command was generated using PuTTY, a free SSH and Telnet client. When using other resources, there might be small graphical variations; however, the basic structure of a PLINK command will be identical

those. Even though the main scientific findings remained supported after appropriate QC, the results of the new analysis deviated strongly enough for the authors to decide to retract the article.

3.1 | Data simulation using HapMap data

To be able to illustrate all analysis steps using realistic genetic data, we simulated a dataset (N = 207) with a binary outcome measure using the publicly available data from the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/plink_format/; Gibbs et al., 2003). For this tutorial, in order to create an ethnically homogenous dataset, we only included Utah residents with ancestry from Northern and Western Europe (CEU). Because of the relatively small sample size of the HapMap data, genetic effect sizes in these simulations were set at values larger than usually observed in genetic studies of complex traits. It is important to note that larger sample sizes (e.g., at least in the order of thousands but likely even tens or hundreds of thousands) will be required to detect genetic risk factors of complex traits. The HapMap data with a simulated phenotypic trait can be found at https://github.com/MareesAT/GWA_tutorial/(1_QC_GWAS.zip).

3.2 | Overview of QC steps

Because of the challenges characterizing GWAS, we aim to illustrate essential QC steps and to provide example scripts. Table 1 provides a summary of seven QC steps and includes recommendations regarding the specific thresholds. However, thresholds may vary according to study specific characteristics. The seven QC steps consist of filtering out of SNPs and individuals based on the following: (1) **individual and**

TABLE 1 Overview of seven QC steps that should be conducted prior to genetic association analysis

Step	Command	Function	Thresholds and explanation
1: Missingness of SNPs and individuals	--geno --mind	Excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed. Excludes individuals who have high rates of genotype missingness. In this step, individual with low genotype calls are removed.	We recommend to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02). Note, SNP filtering should be performed before individual filtering.
2: Sex discrepancy	--check-sex	Checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates.	Can indicate sample mix-ups. If many subjects have this discrepancy, the data should be checked carefully. Males should have an X chromosome homozygosity estimate >0.8 and females should have a value <0.2.
3: Minor allele frequency (MAF)	--maf	Includes only SNPs above the set MAF threshold.	SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on your sample size, larger samples can use lower MAF thresholds. Respectively, for large (N = 100.000) vs. moderate samples (N = 10000), 0.01 and 0.05 are commonly used as MAF threshold.
4: Hardy–Weinberg equilibrium (HWE)	--hwe	Excludes markers which deviate from Hardy–Weinberg equilibrium.	Common indicator of genotyping error, may also indicate evolutionary selection. For binary traits we suggest to exclude: HWE p value <1e−10 in cases and <1e−6 in controls. Less strict case threshold avoids discarding disease-associated SNPs under selection (see online tutorial at https://github.com/MareesAT/GWA_tutorial/). For quantitative traits, we recommend HWE p value <1e-6.
5: Heterozygosity	For an example script see https://github.com/MareesAT/GWA_tutorial/	Excludes individuals with high or low heterozygosity rates	Deviations can indicate sample contamination, inbreeding. We suggest removing individuals who deviate ±3 SD from the samples' heterozygosity rate mean.
6: Relatedness	--genome --min	Calculates identity by descent (IBD) of all sample pairs. Sets threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, pi-hat >0.2 (i.e., second degree relatives) can be detected.	Use independent SNPs (**pruning**) for this analysis and limit it to autosomal chromosomes only. Cryptic relatedness can interfere with the association analysis. If you have a family-based sample (e.g., parent-offspring), you do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population based sample we suggest to use a pi-hat threshold of 0.2, which in line with the literature (Anderson et al., 2010; Guo et al., 2014).
7: Population stratification	--genome --cluster --mds-plot k	Calculates identity by descent (IBD) of all sample pairs. Produces a k-dimensional representation of any substructure in the data, based on IBS.	Use independent SNPs (**pruning**) for this analysis and limit it to autosomal chromosomes only. K is the number of dimensions, which needs to be defined (typically 10). This is an important step of the QC that consists of multiple proceedings but for reasons of completeness we briefly refer to this step in the table. This step will be described in more detail in section "controlling for population stratification."

SNP missingness, (2) inconsistencies in assigned and genetic sex of subjects (see **sex discrepancy**), (3) **minor allele frequency (MAF)**, (4) deviations from **Hardy–Weinberg equilibrium (HWE)**, (5) **heterozygosity** rate, (6) **relatedness**, and (7) ethnic outliers (see **population stratification**).

Box 1:

Clumping: This is a procedure in which only the most significant SNP (i.e., lowest *p* value) in each LD block is identified and selected for further analyses. This reduces the correlation between the remaining SNPs, while retaining SNPs with the strongest statistical evidence.

Co-heritability: This is a measure of the genetic relationship between disorders. The SNP-based co-heritability is the proportion of covariance between disorder pairs (e.g., schizophrenia and bipolar disorder) that is explained by SNPs.

Gene: This is a sequence of nucleotides in the DNA that codes for a molecule (e.g., a protein)

Heterozygosity: This is the carrying of two different alleles of a specific SNP. The heterozygosity rate of an individual is the proportion of heterozygous genotypes. High levels of heterozygosity within an individual might be an indication of low sample quality whereas low levels of heterozygosity may be due to inbreeding.

Individual-level missingness: This is the number of SNPs that is missing for a specific individual. High levels of missingness can be an indication of poor DNA quality or technical problems.

Linkage disequilibrium (LD): This is a measure of non-random association between alleles at different loci at the same chromosome in a given population. SNPs are in LD when the frequency of association of their alleles is higher than expected under random assortment. LD concerns patterns of correlations between SNPs.

Minor allele frequency (MAF): This is the frequency of the least often occurring allele at a specific location. Most studies are underpowered to detect associations with SNPs with a low MAF and therefore exclude these SNPs.

Population stratification: This is the presence of multiple subpopulations (e.g., individuals with different ethnic background) in a study. Because allele frequencies can differ between subpopulations, population stratification can lead to false positive associations and/or mask true associations. An excellent example of this is the chopstick gene, where a SNP, due to population stratification, accounted for nearly half of the variance in the capacity to eat with chopsticks (Hamer & Sirota, 2000).

Pruning: This is a method to select a subset of markers that are in approximate linkage equilibrium. In PLINK, this method uses the strength of LD between SNPs within a specific window (region) of the chromosome and selects only SNPs that are approximately uncorrelated, based on a user-specified threshold of LD. In contrast to clumping, pruning does not take the *p* value of a SNP into account.

Relatedness: This indicates how strongly a pair of individuals is genetically related. A conventional GWAS assumes that all subjects are unrelated (i.e., no pair of individuals is more closely related than second-degree relatives). Without appropriate correction, the inclusion of relatives could lead to biased estimations of standard errors of SNP effect sizes. Note that specific tools for analysing family data have been developed.

Sex discrepancy: This is the difference between the assigned sex and the sex determined based on the genotype. A discrepancy likely points to sample mix-ups in the lab. Note, this test can only be conducted when SNPs on the sex chromosomes (X and Y) have been assessed.

Single nucleotide polymorphism (SNP): This is a variation in a single nucleotide (i.e., A, C, G, or T) that occurs at a specific position in the genome. A SNP usually exists as two different forms (e.g., A vs. T). These different forms are called alleles. A SNP with two alleles has three different genotypes (e.g., AA, AT, and TT).

SNP-heritability: This is the fraction of phenotypic variance of a trait explained by all SNPs in the analysis.

SNP-level missingness: This is the number of individuals in the sample for whom information on a specific SNP is missing. SNPs with a high level of missingness can potentially lead to bias.

Summary statistics: These are the results obtained after conducting a GWAS, including information on chromosome number, position of the SNP, SNP(rs)-identifier, MAF, effect size (odds ratio/beta), standard error, and *p* value. Summary statistics of GWAS are often freely accessible or shared between researchers.

The Hardy–Weinberg (dis)equilibrium (HWE) law: This concerns the relation between the allele and genotype frequencies. It assumes an indefinitely large population, with no selection, mutation, or migration. The law states that the genotype and the allele frequencies are constant over generations. Violation of the HWE law indicates that genotype frequencies are significantly different from expectations (e.g., if the frequency of allele A = 0.20 and the frequency of allele T = 0.80; the expected frequency of genotype AT is 2*0.2*0.8 = 0.32) and the observed frequency should not be significantly different. In GWAS, it is generally assumed that deviations from HWE are the result of genotyping errors. The HWE thresholds in cases are often less stringent than those in controls, as the violation of the HWE law in cases can be indicative of true genetic association with disease risk.

Hands-on experience with the performance of QC Steps 1–7 can be obtained by following all steps outlined in our online tutorial at https://github.com/MareesAT/GWA_tutorial/ (1_QC_GWAS.zip + 2_Population_stratification.zip). It offers scripts for data QC and visualization of potential sources of bias. These scripts perform QC on the CEU group of the HapMap data but can be applied to other datasets with the exception of family-based datasets and datasets involving multiple distinct ethnic groups. Generally, if a sample includes multiple ethnic groups (e.g., Africans, Asians, and Europeans), it is recommended to perform tests of association in each of the ethnic groups separately and to use appropriate methods, such as meta-analysis (Willer, Li, & Abecasis, 2010), to combine the results. If your sample includes subjects from a single ethnic group, the remaining population stratification can be corrected for by the methods discussed below.

4 | CONTROLLING FOR POPULATION STRATIFICATION

An important source of systematic bias in GWAS is population stratification, as explained in Box 1. It has been shown that even subtle degrees of population stratification within a single ethnic population can exist (Abdellaoui et al., 2013; Francioli et al., 2014). Therefore, testing and controlling for the presence of population stratification is an essential QC step.

There are several methods to correct for population stratification (Price, Zaitlen, Reich, & Patterson, 2010). In this tutorial, we illustrate a method that is incorporated in PLINK: the multidimensional scaling (MDS) approach. This method calculates the genome-wide average proportion of alleles shared between any pair of individuals within the sample to generate quantitative indices (components) of the genetic variation for each individual. The individual component scores can be plotted to explore whether there are groups of individuals that are genetically more similar to each other than expected. For example, in a genetic study including subjects from Asia and Europe, MDS analysis would reveal that Asians are genetically more similar to each other than to Europeans. To investigate for which individuals the generated component scores deviate from the samples target population, plotting of the scores of the sample under investigation and a population of known ethnic structure (e.g., HapMap/1KG data) is helpful: This step is called anchoring. This enables the researcher to obtain ethnic information on their sample and to determine possible ethnic outliers. A script is provided at https://github.com/MareesAT/GWA_tutorial/ (2_Population_stratification.zip) to perform MDS on your own data anchored by data of the 1KG project (http://www.1000genomes.org/).

Figure 3 illustrates an example of such an analysis. Individuals who are outliers based on the MDS analysis should be removed from further analyses. After the exclusion of these individuals, a new MDS analysis must be conducted, and its main components need to be used as covariates in the association tests in order to correct for any remaining population stratification within the population. How many components need to be included depends on the population structure and the sample size, but the inclusion of up to 10 components is generally accepted within the psychiatric genetics community.

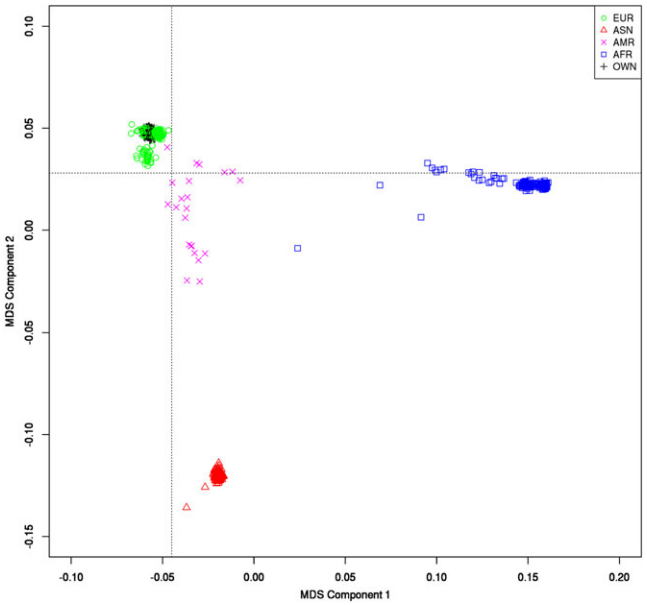

FIGURE 3 Multidimensional scaling (MDS) plot of 1KG against the CEU of the HapMap data (which could be seen as your "own" data in this example, as it is being used in the online tutorial at https://github.com/MareesAT/GWA_tutorial/). The black crosses (+ = "OWN") in the upper left part represent the first two MDS components of the individuals in the HapMap sample (the colored symbols represent the 1KG data (○ = European; □ = African; X = Ad Mixed American; △ = Asian). The MDS components representing the European samples (○) are located in the upper left, the African samples (□) are located in the upper right, the Ad Mixed American samples (X) are located near the intersection point of the dashed lines, the Asian components (△) are located in the lower left part

5 | STATISTICAL TESTS OF ASSOCIATION FOR BINARY AND QUANTITATIVE TRAITS

After QC and calculation of MDS components, the data are ready for subsequent association tests. Depending on the expected genetic model of the trait or disease of interest and the nature of the phenotypic trait studied, the appropriate statistical test can be selected. In the accompanying tutorial, we provide scripts for various types of association that are suitable for binary traits (e.g., alcohol dependent patients vs. healthy controls) or quantitative traits (e.g., the number of alcoholic beverages consumed per week).

PLINK offers one degree of freedom (1 df) allelic tests in which the trait value, or the log-odds of a binary trait, increases or decreases linearly as a function of the number of risk alleles (minor allele [a] vs. major allele [A]). In addition, non-additive tests are available, for instance, the genotypic association test (2 df: aa vs. Aa vs AA), the dominant gene action test (1 df: [aa & Aa] vs AA), and the recessive gene action test, (1 df: aa vs [Aa & AA]). However, non-additive tests are not widely applied, because the statistical power to detect non-additivity is low in practice (Lettre, Lange, & Hirschhorn, 2007; McCarthy et al., 2008). More complex analyses (e.g., Cox regression analysis and cure models) (Stringer, Denys, Kahn, & Derks, 2016) can be performed by using R-based "plug-in" functions in PLINK.

Example scripts for the association analyses described below are located at https://github.com/MareesAT/GWA_tutorial/ (3_Association_GWAS.zip).

5.1 | Binary outcome measure

Within PLINK, the association between SNPs and a binary outcome (value 1 = unaffected and value 2 = affected; 0 and −9 represent missing; the preceding represents the default options in PLINK and can be changed) can be tested with the options `--assoc` or `--logistic`. The `--assoc` option in PLINK performs a X^2 test of association that does not allow the inclusion of covariates. With the `--logistic` option, a logistic regression analysis will be performed which allows the inclusion of covariates. The `--logistic` option is more flexible than the `--assoc` option, yet it comes at the price of increased computational time.

5.2 | Quantitative outcome measure

Within PLINK, the association between SNPs and quantitative outcome measures can be tested with the options `--assoc` and `--linear`. When PLINK detects a quantitative outcome measure (i.e., values other than 1, 2, 0, or missing), the `--assoc` option will automatically treat it as such by performing an asymptotic version of the usual Student's *t* test to compare two means. This option does not allow the use of covariates. The `--linear` option in PLINK performs a linear regression analysis with each individual SNP as a predictor. Similar to the `--logistic` option, the `--linear` option enables the use of covariates and is somewhat slower than the `--assoc` option.

5.3 | Correction for multiple testing

Modern genotyping arrays can genotype up to 4 million markers concurrently, which generates a large number of tests, and thus, a considerable multiple testing burden. SNP imputation may further increase the number of tested associations. Various simulations have indicated that the widely used genome-wide significance threshold of 5×10^{-8} for studies on European populations adequately controls for the number of independent SNPs in the entire genome, regardless of the actual SNP density of the study (Dudbridge & Gusnanto, 2008). When testing African populations, more stringent thresholds are required due to the greater genetic diversity among those individuals (probably close to 1.0×10^{-8}; Hoggart, Clark, De Lorio, Whittaker, & Balding, 2008).

Three widely applied alternatives for determining genome-wide significance are the use of Bonferroni correction, Benjamini–Hochberg false discovery rate (FDR), and permutation testing. The Bonferroni correction, which aims to control the probability of having at least one false positive finding, calculates the adjusted *p* value threshold with the formula $0.05/n$, with *n* being the number of SNPs tested. However, as stated previously, many SNPs are correlated, due to **Linkage Disequilibrium (LD)** and are thus by definition not independent. Therefore, this method is often too conservative and leads to an increase in the proportion of false negative findings.

FDR controls the expected proportion of false positives among all signals with an FDR value below a fixed threshold, assuming that SNPs are independent (Benjamini & Hochberg, 1995). This method is less conservative than Bonferroni correction. It should be noted that controlling for FDR does not imply any notion of statistical significance; it is merely a method to minimize the expected proportion of false positives, for example, for follow-up analyses. Moreover, this method has its own limitation as SNPs and thus *p* values are not independent whereas this is an assumption of the FDR method (Benjamini & Hochberg, 1995). To easily apply Bonferroni and FDR correction, PLINK offers the option `--adjust` that generates output in which the unadjusted *p* value is displayed, along with *p* values corrected with various multiple testing correction methods.

Finally, permutation methods can be used to deal with the multiple testing burden. To calculate permutation-based *p* values, the outcome measure labels are randomly permuted multiple (e.g., 1,000–1,000,000) times which effectively removes any true association between the outcome measure and the genotype. For all permuted data sets, statistical tests are then performed. This provides the empirical distribution of the test-statistic and the *p* values under the null hypothesis of no association. The original test statistic or *p* value obtained from the observed data is subsequently compared to the empirical distribution of *p* values to determine an empirically adjusted *p* value. To use this method, the two PLINK options `--assoc` and `--mperm` can be combined to generate two *p* values: EMP1, the empirical *p* value (uncorrected), and EMP2, the empirical *p* value corrected for multiple testing. This procedure is computationally intensive, especially if many permutations are required, which is necessary to calculate very small *p* values accurately (North, Curtis, & Sham, 2003).

6 | PRS ANALYSIS

6.1 | Computing a PRS

Single variant association analysis has been the primary method in GWAS but requires very large sample sizes to detect more than a handful of SNPs for many complex traits (Gratten, Wray, Keller, & Visscher, 2014; Visscher, Brown, McCarthy, & Yang, 2012). In contrast, PRS analysis does not aim to identify individual SNPs but instead aggregates genetic risk across the genome in a single individual polygenic score for a trait of interest (Purcell et al., 2009; see Figure 4 for a simplified example). In this approach, a large discovery sample is required to reliably determine how much each SNP is expected to contribute to the polygenic score ("weights") of a specific trait. Subsequently, in an independent target sample, which can be more modest in size (Dudbridge, 2013), polygenic scores can be calculated based on genetic DNA profiles and these weights (see below for details on the calculations). As a rule of thumb, a target sample around 2,000 subjects provides sufficient power to detect a significant proportion of variance explained. Furthermore, the discovery and target samples should have the same number of subjects until the target sample includes 2,000 subjects. If more samples are available, additional subjects should be included in the discovery sample to maximize the accuracy of the estimation of the effect sizes (Dudbridge, 2013). Although PRS is not powerful enough to predict disease risk on the individual level (Wray et al., 2013), it has been successfully used to show significant associations both within and across traits. For example, a PRS analysis of schizophrenia showed for the first time that an aggregate measure of the genetic risk to develop schizophrenia, estimated based on the effects of common SNPs (from the discovery sample) that showed nominally significant associations with disease risk, was significantly associated with schizophrenia risk in an independent (target)

Discovery GWAS

	Weight*	Risk Allele
SNP1	0.2	A
SNP2	-0.3	C
SNP3	0.1	G

Individual	Alleles SNP1	Alleles SNP2	Alleles SNP3
1	AT	AA	CG
2	AA	CA	GG
3	TT	AC	CG
4	TT	AA	GG
5	TA	CA	GC
6	AT	CA	CG
7	AA	AA	GG
8	AA	CC	CG
9	TA	CC	GC
10	AT	AA	CG

PRS:

Individual	SNP 1	SNP 2	SNP 3	**PRS**
1	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**
2	0.2+0.2	-0.3+0.0	0.1+0.1	**0.3**
3	0.0+0.0	0.0-0.3	0.0+0.1	**-0.2**
4	0.0+0.0	0.0+0.0	0.1+0.1	**0.2**
5	0.0+0.2	-0.3+0.0	0.1+0.0	**0.0**
6	0.2+0.0	-0.3+0.0	0.0+0.1	**0.0**
7	0.2+0.2	0.0+0.0	+0.1+0.1	**0.6**
8	0.2+0.2	-0.3-0.3	0.0+0.1	**-0.1**
9	0.0+0.2	-0.3-0.3	0.1+0.0	**-0.3**
10	0.2+0.0	0.0+0.0	0.0+0.1	**0.3**

FIGURE 4 Working example of three single nucleotide polymorphisms (SNPs) aggregated into a single individual polygenic risk score (PRS). *The weight is either the beta or the log of the odds-ratio, depending on whether a continuous or binary trait is analysed

sample. The significant association was found despite the fact that the available sample sizes were too small to detect genome-wide significant SNPs (Purcell et al., 2009). In addition, GWAS for schizophrenia (the discovery sample) has been used to significantly predict the risk in target samples with various phenotypes, such as bipolar disorder, level of creativity, and even risk of immune disorders (Power et al., 2015; Purcell et al., 2009; Stringer et al., 2014; Wray et al., 2013).

To conduct PRS analysis, trait-specific weights (beta's for continuous traits and the log of the odds ratios for binary traits) are obtained from a discovery GWAS. In the target sample, a PRS is calculated for each individual based on the weighted sum of the number of risk alleles that he or she carries multiplied by the trait-specific weights. For many complex traits, SNP effect sizes are publicly available (e.g., see https://www.med.unc.edu/pgc/downloads).

Although in principle all common SNPs could be used in a PRS analysis, it is customary to first clump (see **clumping**) the GWAS results before computing risk scores. *p* value thresholds are typically used to remove SNPs that show little or no statistical evidence for association (e.g., only keep SNPs with *p* values <0.5 or <0.1. Usually, multiple PRS analyses will be performed, with varying thresholds for the *p* values.

6.2 | Conducting polygenic risk prediction analyses

Once PRS have been calculated for all subjects in the target sample, the scores can be used in a (logistic) regression analysis to predict any trait that is expected to show genetic overlap with the trait of interest. The prediction accuracy can be expressed with the (pseudo-) R^2 measure of the regression analysis. It is important to include at least a few MDS components as covariates in the regression analysis to control for population stratification. To estimate how much variation is explained by the PRS, the R^2 of a model that includes only the covariates (e.g., MDS components) and the R^2 of a model that includes covariates + PRS will be compared. The increase in R^2 due to the PRS indicates the increase in prediction accuracy explained by genetic risk factors.

The prediction accuracy of PRS depends mostly on the **(co-)heritability** of the analysed traits, the number of SNPs, and the size of the discovery sample. The size of the target sample only affects the reliability of R^2 and typically a few thousand of subjects in the target sample are sufficient to achieve a significant R^2 if the (co-)heritability of the trait(s) of interest and the sample size of the discovery sample used are sufficiently large. For an R script to perform power calculations for your own PRS analysis, we refer to the POLYGENE script on https://sites.google.com/site/fdudbridge/software (Dudbridge, 2013).

A convenient program to perform PRS analysis is PRSice (see http://prsice.info; Euesden, Lewis, & O'Reilly, 2015). It takes care of clumping, *p* value thresholds, MDS components, and plots attractive graphs. We refer to https://github.com/MareesAT/GWA_tutorial/(4_PRS.doc) for a tutorial on how to perform your own PRS analysis using PRSice. Other programs for the application of PRS are, for example, PLINK (`--score`) and LDpred (Purcell et al., 2007; Vilhjalmsson et al., 2015).

7 | CONCLUSION

A basic understanding of the theory behind genetic analysis (e.g., GWAS and PRS), the essential QC steps, and the use of appropriate software and methods, along with practical experience are imperative to be able to conduct a genetic study with reliable and reproducible results. This tutorial highlights important concepts to successfully conduct a GWAS and PRS analysis. We presented a tutorial based on commonly used, open-source, freely available software tools, that are accessibly for novice users. In addition, we made scripts and a simulated data set available to provide hands-on practice at https://github.com/MareesAT/GWA_tutorial/.

As a GWAS is usually undertaken to increase our understanding of the biological mechanisms that contribute to disease risk, a GWAS will usually be followed up by post-GWAS analyses. Valuable insights can be acquired by using tools and resources, which enable the researcher to interpret the association results from a functional or biological perspective. GTEx provides information on the association between SNPs and gene expression (Ardlie et al., 2015). Ensembl (Birney et al., 2004) and FUMA (Watanabe, Taskesen, van Bochoven, & Posthuma, 2017) are commonly used tools for functional annotation. In addition, methods that provide important insights into the genetic architecture of the psychiatric trait or disease under study are freely available. For

example, GCTA (Yang, Lee, Goddard, & Visscher, 2011) and LD score regression analysis (Bulik-Sullivan et al., 2015) have been applied to estimate **SNP-based heritability**. Gene-based tests, which consider the association between a phenotypic trait and multiple SNPs within a **gene**, (e.g., de Leeuw, Neale, Heskes, & Posthuma, 2016) and pathway/gene-set analyses (de Leeuw et al., 2016) have increased our insight into the biological pathways of psychiatric disorders. It should be noted that many of the aforementioned methods can be applied using **summary statistics**. It is beyond the scope of this paper to discuss all available post-GWAS tools and resources in detail. For in-depth information on post-GWAS analyses, we refer to an excellent article by Reed and colleagues (Reed et al., 2015).

ACKNOWLEDGEMENTS

A. T. M. and E. M. D. are supported by the Foundation Volksbond Rotterdam. F. V. is supported by the Investissement d'Avenir programme managed by the ANR under reference ANR-11-IDEX-0004-02.

DECLARATION OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

ORCID

Andries T. Marees http://orcid.org/0000-0002-9781-7492

REFERENCES

Abdellaoui, A., Hottenga, J. J., Xiao, X. J., Scheet, P., Ehli, E. A., Davies, G. E., … Boomsma, D. I. (2013). Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior Genetics*, *43*(6), 455–467. https://doi.org/10.1007/s10519-013-9610-1

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. https://doi.org/10.1038/nprot.2010.116

Ardlie, K. G., DeLuca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Consortium, G. (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). Gen ABEL: An R library for genome-wide association analysis. *Bioinformatics*, *23*(10), 1294–1296. https://doi.org/10.1093/bioinformatics/btm108

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B Methodological*, *57*(1), 289–300.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., … Clamp, M. (2004). An overview of Ensembl. *Genome Research*, *14*(5), 925–928. https://doi.org/10.1101/gr.1860604

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., … Nervos, G. C. A. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*(11), 1236–1241. https://doi.org/10.1038/ng.3406

Chen, M. H., & Yang, Q. (2010). GWAF: An R package for genome-wide association analyses with family data. *Bioinformatics*, *26*(4), 580–581. https://doi.org/10.1093/bioinformatics/btp710

de Leeuw, C. A., Neale, B. M., Heskes, T., & Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nature Reviews Genetics*, *17*(6), 353–364. https://doi.org/10.1038/nrg.2016.29

Derks, E. M., Vorstman, J. A. S., Ripke, S., Kahn, R. S., Ophoff, R. A., & Con, S. P. G. (2012). Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: A polygenic risk score analysis. *Plos One*, *7*(6). Doi: ARTN e37852 https://doi.org/10.1371/journal.pone.0037852), e37852.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, *9*(3). Doi: ARTN e1003348 https://doi.org/10.1371/journal.pgen.1003348

Dudbridge, F. (2016). Polygenic epidemiology. *Genetic Epidemiology*, *40*(4), 268–272. https://doi.org/10.1002/gepi.21966

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics*, *31*(9), 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

Francioli, L. C., Menelaou, A., Pulit, S. L., Van Dijk, F., Palamara, P. F., Elbers, C. C., … Study, L. C. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825. https://doi.org/10.1038/ng.3021

Gelernter, J., Sherva, R., Koesterer, R., Almasy, L., Zhao, H., Kranzler, H. R., & Farrer, L. (2014). Genome-wide association study of cocaine dependence and related traits: FAM53B identified as a risk gene. *Molecular Psychiatry*, *19*(6), 717–723. https://doi.org/10.1038/mp.2013.99

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., … Consortium, I. H. (2003). The international hap map project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, *13*(2), 135–145. https://doi.org/10.1038/nrg3118

Gratten, J., Wray, N. R., Keller, M. C., & Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience*, *17*(6), 782–790. https://doi.org/10.1038/nn.3708

Guo, Y., He, J., Zhao, S. L., Wu, H., Zhong, X., Sheng, Q. H., … Long, J. R. (2014). Illumina human exome genotyping array clustering and quality control. *Nature Protocols*, *9*(11), 2643–2662. https://doi.org/10.1038/nprot.2014.174

Hamer, D., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry*, *5*(1), 11–13.

Hoggart, C. J., Clark, T. G., De Lorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, *32*(2), 179–185. https://doi.org/10.1002/gepi.20292

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, *31*(4), 358–362. https://doi.org/10.1002/gepi.20217

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913. https://doi.org/10.1038/ng2088

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

North, B. V., Curtis, D., & Sham, P. C. (2003). A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics*, *72*(2), 498–499. https://doi.org/10.1086/346173

Ott, J., Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, *12*(7), 465–474. https://doi.org/10.1038/nrg2989

Power, R. A., Steinberg, S., Bjornsdottir, G., Rietveld, C. A., Abdellaoui, A., Nivard, M. M., … Stefansson, K. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience*, *18*(7), 953–955. https://doi.org/10.1038/nn.4040

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459–463. https://doi.org/10.1038/nrg2813

Purcell, S. M., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Scolnick, E. M. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. https://doi.org/10.1038/nature08185

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, *34*(28), 3769–3792. https://doi.org/10.1002/sim.6605

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., … Consor, W. T. C.-C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421. https://doi.org/10.1038/nature13595

Ruderfer, D. M., Fanous, A. H., Ripke, S., McQuillin, A., Amdur, R. L., Gejman, P. V., … Psychia, C.-D. W. G. (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry*, *19*(9), 1017–1024. https://doi.org/10.1038/mp.2013.138

Sebastiani, P., Bae, H., Sun, F. G. X., Andersen, S. L., Daw, E. W., Malovini, A., … Perls, T. T. (2013). Meta-analysis of genetic variants associated with human exceptional longevity. *Aging-Us*, *5*(9), 653–661.

Sebastiani, P., Solovieff, N., DeWan, A. T., Walsh, K. M., Puca, A., Hartley, S. W., … Perls, T. T. (2012). Genetic signatures of exceptional longevity in humans. *Plos One*, *7*(1). Doi: ARTN e29848 https://doi.org/10.1371/journal.pone.0029848), e29848.

Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Andersen, S., … Perls, T. T. (2010). Genetic signatures of exceptional longevity in humans. *Science*, *2010*. https://doi.org/10.1126/science.1190532

Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Andersen, S., … Perls, T. T. (2011). Editorial expression of concern (retraction of vol 330, pg 912, 2010). *Science*, *333*(6041), 404–404.

Shieh, Y., Hu, D. L., Ma, L., Huntsman, S., Gard, C. C., Leung, J. W. T., … Ziv, E. (2016). Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Research and Treatment*, *159*(3), 513–525. https://doi.org/10.1007/s10549-016-3953-2

Smoller, J. W. (2013). Cross Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis (vol 381, pg 1371, 2013). *Lancet*, *381*(9875), 1360–1360.

Stringer, S., Cerrone, K. C., van den Brink, W., van den Berg, J. F., Denys, D., Kahn, R. S., & Derks, E. M. (2015). A guide on gene prioritization in studies of psychiatric disorders. *International Journal of Methods in Psychiatric Research*, *24*(4), 245–256. https://doi.org/10.1002/mpr.1482

Stringer, S., Denys, D., Kahn, R. S., & Derks, E. M. (2016). What cure models can teach us about genome-wide survival analysis. *Behavior Genetics*, *46*(2), 269–280. https://doi.org/10.1007/s10519-015-9764-0

Stringer, S., Kahn, R. S., de Witte, L. D., Ophoff, R. A., & Derks, E. M. (2014). Genetic liability for schizophrenia predicts risk of immune disorders. *Schizophrenia Research*, *159*(2–3), 347–352. https://doi.org/10.1016/j.schres.2014.09.004

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Disease mechanisms genetic architectures of psychiatric disorders: The emerging picture and its implications. *Nature Reviews Genetics*, *13*(8), 537–551. https://doi.org/10.1038/nrg3240

van Leeuwen, E. M., Kanterakis, A., Deelen, P., Kattenberg, M. V., Slagboom, P. E., de Bakker, P. I. W., … Consortium, G. N. (2015). Population-specific genotype imputations using minimac or IMPUTE2. *Nature Protocols*, *10*(9), 1285–1296. https://doi.org/10.1038/nprot.2015.077

Vassos, E., Di Forti, M., Coleman, J., Iyegbe, C., Prata, D., Euesden, J., … Breen, G. (2017). An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biological Psychiatry*, *81*(6), 470–477. https://doi.org/10.1016/j.biopsych.2016.06.028

Vilhjalmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., … Inherited, D. B. R. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). FUMA: Functional mapping and annotation of genetic associations. *bio Rxiv*, doi:https://doi.org/10.1101/110023.

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, *14*(7), 507–515. https://doi.org/10.1038/nrg3457

Yang, J. A., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

How to cite this article: Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27:e1608. https://doi.org/10.1002/mpr.1608